Post-Print: Final Draft Post Refereeing
Published Version: Shreeves, Sarah L., Thomas G. Habing, Kat Hagedorn, and Jeffery Young. 2005. Current developments and future trends for the OAI Protocol for Metadata Harvesting. Library Trends 53, no. 4: 576-589.

Current Developments and Future Trends for the OAI Protocol for Metadata Harvesting

Sarah Shreeves

Visiting Assistant Professor of Library Administration

IMLS DCC

University of Illinois at Urbana-Champaign

1301 W. Springfield, Room 52

Urbana, IL 61801

Tel: 217.244.7809

Email: sshreeve@uiuc.edu

Thomas Habing

Research Programmer

Room 155 Grainger Engineering Library

1301 W. Springfield Ave.

Urbana, IL 61801

Tel: 217.244.4425

Email: thabing@uiuc.edu

Post-Print: Final Draft Post Refereeing
Published Version: Shreeves, Sarah L., Thomas G. Habing, Kat Hagedorn, and Jeffery Young. 2005. Current developments and future trends for the OAI Protocol for Metadata Harvesting. Library Trends 53, no. 4: 576-589.

Kat Hagedorn

OAIster/Metadata Harvesting Librarian

DLXS Bibliographic Class Coordinator

DLXS Text Class Collections Co-coordinator

Digital Library Production Service

University of Michigan

Tel: 734.615.7618

Email: khage@umich.edu

Jeffrey Young

Software Architect

Online Computer Library Center (OCLC)

6565 Frantz Rd., 43017 Dublin, OH, USA

Tel: 614.764.4342

Email: jyoung@oclc.org

Abstract

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) has been widely adopted since its initial release in 2001. Initially developed as a means to federate access to diverse e-print archives through metadata harvesting and aggregation,

the protocol has demonstrated its potential usefulness to a broad range of communities. Two years out from the release of the stable production version of the protocol (2.0), there are many interesting developments within the OAI community. Communities of interest have begun to use the protocol to aggregate metadata relative to their needs. The development of a registry of OAI data providers with browsing and searching capabilities as well as accessibility to machine processing is helping to provide a scalable solution to the question of who is providing what via the OAI protocol. Work is progressing on the technical infrastructure for extending the OAI protocol beyond the traditional harvesting structure. However, serious challenges, particularly for service providers, still exist. This paper provides an overview of the current OAI environment and speculates on future directions for the protocol and OAI community.

Brief Author Biographies

Sarah L. Shreeves is Visiting Assistant Professor of Library Administration and Project Coordinator for the University of Illinois IMLS Digital Collections and Content Project. Previously she was a Project Coordinator for the University of Illinois Open Archives Initiative Metadata Harvesting Project funded by the Andrew W. Mellon Foundation. From 1992 until 2001 she was a member of staff at the Massachusetts Institute of Technology Libraries. She has a B.A. in Medieval Studies from Bryn Mawr College, an M.A. in Children's Literature from Simmons College, and an M.S. in Library and Information Science from the University of Illinois.

Thomas Habing is a Research Programmer at the Grainger Engineering Library Information Center at the University of Illinois, Urbana-Champaign (UIUC), where for the past six years he has worked on various digital library projects including UIUC's Mellon-funded OAI Metadata Harvesting Project, UIUC's D-Lib Test Suite project funded by CNRI, and UIUC's Digital Library Initiatives I project funded by NSF. He is currently providing programming support for the University Library's NSF-NSDL and IMLS grant projects and is the developer of the Library's OAI Registry service.

Kat Hagedorn is OAIster / Metadata Harvesting Librarian at the University of Michigan Libraries. She is responsible for the OAIster project, a search gateway for OAI harvested records leading to digital objects, initially Mellon-funded in 2001-2002. She is also responsible for DLXS Bibliographic Class and co-coordinates the processing of Text Class materials. Her previous experience is in information architecture (with the Argus Associates firm) and ontology and taxonomy consulting (with the Food and Agriculture Organization in Rome). She graduated with an undergraduate degree in Biological Sciences from Cornell University and got her MLIS at the University of Michigan in 1996.

Jeff Young is a Software Architect for the OCLC Office of Research. He has worked at OCLC since 1987 and in the Office of Research since 1996. He holds a B.S. (Computer Science) from Ohio State and M.L.S (Beta Phi Mu) from Kent State. Current research interests include web services, interoperability, and authority control. He first got involved with OAI in association with the Networked Digital Library of Theses and

Dissertations, and was a member of the OAI Technical Committee that helped develop the OAI-PMH specification.

Post-Print: Final Draft Post Refereeing
Published Version: Shreeves, Sarah L., Thomas G. Habing, Kat Hagedorn, and Jeffery Young. 2005. Current developments and future trends for the OAI Protocol for Metadata Harvesting. Library Trends 53, no. 4: 576-589.

Current Developments and Future Trends for the OAI Protocol for Metadata Harvesting

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) has been widely adopted since its initial release in 2001. Initially developed as a means to federate access to diverse e-print archives through the metadata harvesting (Lagoze & Van de Sompel, 2003), the protocol has demonstrated its potential usefulness to a broad range of communities. According to the Experimental OAI Registry at The University of Illinois Library at Urbana-Champaign (UIUC) (*Experimental OAI Registry*, n.d.), there are currently over 300 active data providers using the production version (2.0) of the protocol from a wide variety of domains and institution types. Developers of both open source and commercial content management systems (such as D-Space and CONTENTdm) are including OAI data provider services as part of their products. Service providers range from large-scale efforts with a wide scope, such as the National Science Digital Library (*NSDL*, n.d.), to small tightly focused community-specific services, such as the Sheet Music Consortium (*Sheet Music Consortium*, n.d.).

This article provides a brief overview of the OAI environment, two years out from the release of the production version of the protocol. We assume a relatively high level of familiarity with how the protocol works and only give a brief overview. We delve into some of the interesting developments within the OAI world, particularly the use of the protocol within specific communities of interest, the development of a comprehensive registry of OAI data providers, and a resolver for OAI identifiers that extends the protocol beyond its traditional use. We also document some of the current challenges for

both data and service providers. We end the paper by noting some of the possible future directions for the OAI protocol and community.

Current Developments in OAI Work

The mission of the Open Archives Initiative, the entity responsible for the protocol, is to "develop and promote interoperability standards that aim to facilitate the efficient dissemination of content." (*Open Archives Initiative*, n.d.) The Protocol for Metadata Harvesting, a tool developed through the OAI, facilitates interoperability between disparate and diverse collections of metadata through a relatively simple protocol based on common standards (XML, HTTP, and Dublin Core). The OAI world is divided into *data providers* or *repositories,* which traditionally make their metadata available through the protocol and *service providers* or *harvesters* who completely or selectively harvest metadata from data providers again through the use of the protocol. (Lagoze & Van de Sompel, 2001) The OAI protocol requires that data providers expose metadata in at least unqualified Dublin Core; however, the use of other metadata schemas is possible and encouraged. The protocol can provide access to parts of the 'invisible web' that are not easily accessible to search engines (such as resources within databases) (Sherman & Price, 2003) and can provide ways for communities of interest to aggregate resources from geographically diffuse collections. The protocol promotes a structure in which data providers can focus on building collections and content, and service providers can focus on building services for these collections and content. While the protocol itself says nothing about what happens to metadata once harvested, usually service providers aggregate, index, and build search/retrieval and other value-added services around the

harvested metadata. It has been now two years since the production version of the protocol was introduced. (Lagoze & Van de Sompel, 2002) Below we discuss just some of the current trends and developments within the OAI community.

Community-/Domain-Specific OAI Services

As mentioned above, the Open Archives Initiative emerged from and was initially designed to meet the needs of the e-print archives community (Warner, 2003). However, it was recognized fairly early in the protocol's development that it could be applicable in a broad range of communities, including, but not limited to libraries, museums, and archives. In fact the implementation guidelines (Lagoze, Van de Sompel, Nelson, & Warner, 2002) are deliberately non-specific so as to provide room for community-specific applications of the protocol. (Lagoze & Van de Sompel, 2003)

The initial push for developing OAI service providers was in part due to the Andrew W. Mellon Foundation grants in 2001 (Waters, 2001). The Foundation issued seven grants to institutions interested in researching the development of service providers. Three institutions developed publicly accessible services predicated on their research: the AmericanSouth.org project at Emory University; the Digital Gateway to Cultural Heritage Materials at the University of Illinois at Urbana-Champaign (UIUC); and the OAIster project at The University of Michigan. Each service had a different focus. The AmericanSouth.org project focused on aggregating content related to the culture and history of the American South while involving scholars in the process of selection and interpretation (Halbert, 2003). The UIUC project aggregated metadata relating to cultural heritage resources including finding aids (Shreeves, Kaczmarek, and Cole, 2003), and the

OAIster project harvested all possible repositories but kept only those records that pointed to actual digital objects (Hagedorn, 2003).

The different foci were indicative of the future progress of service providers. No one service provider can serve the needs of the entire public, hence user-group-specific service providers have become the norm. Many communities have adopted or are in the process of adopting the OAI protocol to help provide federated access to dispersed resources. These communities of interest are significant not only because they have adopted the protocol for a specific domain, but also because they have developed additional standards, tools, and metadata schemas to use along with the OAI protocol--much as the originators of the protocol had hoped. Indeed, these domain- and user-specific services may be the best example of what the OAI protocol has to offer.

We highlight three notable community- or domain-specific services in various stages of development below. For a fuller documentation of community specific service providers and data providers, see the 2003 Digital Library Federation report (Brogan, 2003) and the recent series of profiles of service providers in *Library Hi Tech News*. (McKiernan, 2003a; McKiernan, 2003b; McKiernan, 2004)

Open Language Archives Community The mission of the Open Language Archives Community (OLAC) is to create "a worldwide virtual library of language resources" through development of community-based standards for archiving and interoperability and a "network of interoperable repositories" (*OLAC*, n.d.). OLAC uses the OAI protocol as a means to the latter end. OLAC has extended the protocol to meet the needs for their particular community, specifically through the maintenance of a specialized metadata schema (based loosely on unqualified Dublin Core), data provider tools (including a

range of options for organizations without the technical infrastructure to support full fledged OAI data providers), and service provider tools (Simons & Bird, 2003). Currently OLAC provides access to metadata harvested from 27 data providers through search services hosted at the Linguist List (*Linguist List*, n.d.) and the Linguistic Data Consortium. (*Linguistic Data Consortium*, n.d.) This integration of search services within important community web sites increases the visibility and value of OLAC.

Sheet Music Consortium The Sheet Music Consortium is a group of four academic libraries--UCLA, Johns Hopkins University, Indiana University, and Duke University--which are building a freely available collection of digitized sheet music. Sheet music presents a particular problem for cataloging because of its various elements: cover art, the sheet music itself, the lyrics, etc. (Davison, Requardt & Brancolini, 2003). The Consortium provides standards for using unqualified Dublin Core to describe sheet music and guidelines for implementation of data provider services. The search service allows the creation of 'virtual collections' and allows users to annotate the metadata records (*Sheet Music Consortium*, n.d.). While work on this service is still in progress, the focus on building a service provider based on a specific type of material makes it well worth watching.

National Science Digital Library The National Science Digital Library (NSDL) provides access to the content of collections of science-based learning objects (*NSDL*, n.d.). The OAI protocol is the primary means of aggregating the metadata describing this content,

although other means are used as well (Lagoze et al, 2002). Funded by the National Science Foundation, the NSDL has the broadest vision of the service providers described here in that it is attempting to build and aggregate not just a series of digital collections and content, but services to use these resources and the infrastructure to support both. As such, NSF has invested significant resources to the development of content, services, and infrastructure. The NSDL maintains standards for metadata and guidance for data providers. The NSDL aims for a broad user base (K-12) but its core mission remains to develop this "learning environment and resources network" for science education (Zia, 2001).

Comprehensive OAI Registry of Data Providers

As the OAI community has matured, and especially as the number of OAI repositories and the number of data sets served by those repositories has grown, it has become increasingly difficult for service providers to discover and effectively utilize the myriad repositories. In order to address this difficulty the OAI research group at the University of Illinois Library at Urbana-Champaign (UIUC) has developed a comprehensive, searchable registry of OAI repositories (*Experimental OAI Registry*, n.d.)..

Shortcomings of Existing Registries

There were and continue to be several other registries of OAI repositories such as those maintained by the Open Archives Initiative website (*OAI registered*, n.d.) and OLAC (*OLAC participating*, n.d.). However, nearly all of these suffer from a number of

shortcomings. Probably foremost is that the registries typically maintain very sparse records about the individual repositories, usually nothing other than flat lists of base URLs, possibly including the repository name. Typically there is no search mechanism and fairly limited browse capabilities. An onerous amount of manual snooping using the OAI-PMH verbs directly in a web browser is usually required by potential service providers before they can assess the utility of a specific repository for their needs.

A second shortcoming of the existing registries is completeness. The registries are usually populated by self-registration or maintained to support the specific needs of a unique community, so few of the registries approach a complete list of all available repositories. 'Googling' or following friends or provenance links discovered many new OAI repositories that were not listed in any of the existing registries, even taken as a whole.

Developing the Experimental OAI Registry

In developing OAI service providers for various projects within the UIUC Library, the issues of completeness and discoverability have become more evident. The UIUC research group thus built Experimental OAI Registry to address these problems. Moreover, based on feedback after the first public announcement of the Registry on the OAI-Implementers listserv, the group realized that the Registry could also be utilized to meet various other needs in the OAI community, such as the need for various output formats to support machine processing of the Registry.

Completeness The UIUC research group addressed the completeness issue by employing three different strategies. The first strategy was a simple inventory of existing registries, both formal and informal, that listed different repositories. The second strategy involved following various links that were contained within the OAI responses. The first source of links was the 'friends' container (Lagoze, Van de Sompel, Nelson & Warner, 2002). This container could be included as one of the optional description elements in an OAI 'Identify' response. It allows an OAI repository to list other confederate repositories that may be of interest to a harvester. It is also commonly used by aggregator repositories. The other source of links was the 'provenance' container. (Lagoze, Van de Sompel, Nelson & Warner, 2002) This container could be included as one of the optional 'about' elements of an OAI record. The provenance container stores data about the original source of a record that has been aggregated into a different repository. Using 'friends' and 'provenance' it was possible to recursively crawl webs of related OAI repositories. The registry maintains this linking information about each repository to produce a network graphic.The third strategy involved using the Google™ SOAP-based Web toolkit (*Google Web APIs*, n.d.). Using this toolkit the research group was able to programmatically search the Google™ web indexes to find OAI repositories. The group developed a number of search strategies, from using OAI related keywords such as 'OAI' or 'Open Archives,' to using special Google™ keywords such as 'allinurl:verb=Identify' which will find web sites that contain the string 'verb=Identify' in their URL. This latter strategy proved the most successful. Once a candidate base URL is discovered it is tested

to determine whether it can respond to the OAI 'verb=Identify' request. If it responds, it is assumed to be a valid OAI repository and it is added to the registry.

Finally, requests to manually add repositories to the registry are accepted. In the future, self registration should become an automated procedure.

Searchable and Browsable The second major objective was to make it possible to search for OAI repositories using various criteria, and browse through different views of the registry, but without any manual cataloging of the various OAI repositories. To accomplish this the research group developed processes to automatically harvest and index various data from each repository. Essentially, a specialized harvest of each repository is performed. This harvest collects data from the Identify, ListSets, and ListMetadataFormats responses, supplying these data to various tables and fields in a relational database. In addition, sample records from each OAI repository are collected for each combination of set and metadataPrefix supported by the provider. These data are also added to the relational database. Once these data are indexed, including the full-text of each response, various searches and views of the registry are possible.

The primary supported search is for keywords appearing in the various OAI responses, namely Identify, ListSets, and the sample records. A key observation resulting from our search system is that repositories, including rich collection level metadata either in the optional Identify description containers or the optional ListSets setDescription containers will fare better in terms of discoverability. This suggests the desirability of broader use of collection-level metadata by the OAI community.

Amenable to Machine Processing The third major goal was to expose the registry's data in ways that were useful for machine processing. The most obvious way to make the registry accessible for machine processing was by making it an OAI repository itself. Thus, basic Dublin Core records about each OAI repository contained in the registry can be harvested via the OAI-PMH. The ERRoL service, described below, is an example of an application that utilizes the OAI-PMH interface to the registry. In the future, additional metadata formats might be harvestable as well, such as the ZeeRex format used by the SRW/U protocol (*ZeeRex*, n.d.). In addition, the registry is also an RDF Site Summary (RSS) news feed provider. Using RSS a person can monitor the registry for new or modified repository records. The RSS feed is available off of the registry Web site (*Experimental OAI Registry*, n.d.). There are also a number of ways to export repository records from the registry. Any list of repositories resulting from a search or a browsable view can be exported using the XML schema of the 'friends' description container.

Work is also progressing on a 'harvest bag' feature. This would allow a user to accumulate a custom list of repositories, including sets and metadata formats, that they could export in a standard XML schema. This would be similar to the 'book bag' feature of other digital library portals, which allow users to save and export lists of bibliographic citations. The vision is that the 'harvest bag' list could then be imported into harvesting software to initiate a harvest of the selected sites.

In addition, the research group is working on a SRW/U search service for the registry (*SRW*, n.d.). This would allow SRW/U clients to search the registry in a manner similar to that provided by the web forms search interface. The record formats available

via the SRW/U interface would be the same as those available via the registry's OAI provider.

Future Work

While the registry is now fully operational, there remain a number of improvements the group would like to make to increase its usefulness. Following, in no order, are some plans for future enhancements to the registry:

- Enhance the collection-level description of the repositories to enable better search and discover. This might include both manual cataloging and the application of automated classification algorithms to the repository's records.
- Provide more automated maintenance of the registry, including the ability of OAI data providers to securely add or modify their repository's records in the registry, including collection-level descriptive data.
- Improve the automated discovery of new repositories, such as automatically running the Google™ SOAP-based harvester.
- Delegate the creation and maintenance of virtual collections of repositories, including collection-level metadata.
- Improve the view of search results, especially the context of the search hit. The current system does not identify the context of a search hit, which could be the Identify or ListSets responses or the sample records.

Extensible Repository Resource Locators (ERRoLs)

As mentioned above, according to the conventional model of OAI, the world is divided into data providers and service providers. As it happens, though, a few simple tricks with stylesheets and HTTP redirects allow an OAI repository to stand alone as an independent web application. Early examples of this were created by enhancing individual repositories as discussed elsewhere (Van de Sompel, Young, & Hickey, 2003). Frustration with changing the OAI world one repository at a time, though, led to the development of the ERRoL resolution service (*ERROLs*, n.d.) that automatically extends these same features and more to any OAI repository in the UIUC registry.

ERRoLs are "Cool URLs" (Berners-Lee, 1998) to content and services related to information in an OAI repository. In essence, the ERRoL service is a resolver for oai-identifiers. In its simplest form, the oai-identifier for an item (such as "oai:lcoa1.loc.gov:loc.pnp/cph.3b37282") can be resolved by appending it to the end of the ERRoL service URL "http://errol.oclc.org/" as in "http://errol.oclc.org/oai:lcoa1.loc.gov:loc.pnp/cph.3b37282". The ERRoL service begins the resolution process by parsing the repository-identifier ("lcoa1.loc.gov") from the URL and using it to obtain the official OAI baseURL from the UIUC registry. With this, the ERRoL service constructs a standard OAI GetRecord (oai_dc) request to the home repository, which is what the client sees in response.

As a resolution result, however, an XML OAI GetRecord response is of marginal interest at best. Fortunately, appending various extensions to the basic URL form can produce different kinds of results. For example, if we want this same oai_dc record stripped from the OAI GetRecord wrapper, we can append the "oai_dc" metadataPrefix

to the URL, as in "http://errol.oclc.org/oai:lcoa1.loc.gov:loc.pnp/cph.3b37282.oai_dc." This home repository can also supply a "marcxml" record for this same oai-identifier, which can be obtained by appending a ".marcxml" extension, as in "http://errol.oclc.org/oai:lcoa1.loc.gov:loc.pnp/cph.3b37282.marc21". Any metadataPrefix available for this item can be added as an extension. This ability to strip a record from its OAI GetRecord wrapper becomes particularly interesting when OAI repositories contain XML *content,* beyond metadata. Here are examples for a repository that can disseminate XHTML (metadataPrefix=xhtml), XSL Stylesheets (metadataPrefix=xsl), and XML Schemas (metadataPrefix=xsd) respectively:

- http://errol.oclc.org/oai:xmlregistry.oclc.org:xoai/xoaiharvester.xhtml
- http://errol.oclc.org/oai:xmlregistry.oclc.org:xoai/xoaiharvester.xsl
- http://errol.oclc.org/oai:xmlregistry.oclc.org:xoai/config.xsd

Keep in mind that the ERRoL service is stripping these XML documents from OAI GetRecord responses that it retrieves from the home repository. Each shares the same oai-identifier as the oai_dc metadata record that describes it, which, as explained above, can be obtained by changing the extension to "oai_dc". Having content and metadata in such close proximity makes it easy to build lightweight, interactive, self-descriptive, content-based, automated systems using XSLT and other thin clients.

These examples demonstrate that ERRoLs are a simple mechanism for accessing various manifestations of OAI data, but it cannot be said that they elevate an OAI repository to the level of a human-interactive web application yet. But just as ERRoLs

transformed standard OAI responses into other forms in the examples above, they can just as easily transform them into HTML using the ".html" extension, as in "http://errol.oclc.org/oai:lcoa1.loc.gov:loc.pnp/cph.3b37282.html." The ".html" extension, as well as others, not only works at the item level with oai-identifiers, but also at the repository level with *repository*-identifiers. In the case of repository-identifier "lcoa1.loc.gov." URL patterns like "http://errol.oclc.org/lcoa1.loc.gov.html" are possible. Furthermore, standard OAI parameters can be appended to this URL to produce HTML renderings of all the OAI-PMH responses, as in "http://errol.oclc.org/xmlregistry.oclc.org.html?verb=ListRecords&metadataPrefix=oai_dc&set=XSLStylesheets".

ERRoLs work with any OAI repository that has a unique repository-identifier registered at the UIUC Experimental OAI Registry. In the case of the ".html" extension, the repository displays integrate identity and branding information gleaned from the repository's 'Identify' response, but otherwise the repositories share the same look and feel. It is possible, however, for individual repositories to instruct the ERRoL service to use an alternate stylesheet by inserting a <description> element in their 'Identify' response. Thus, the GSAFD Thesaurus repository (*GSAFD Thesaurus*, n.d.) looks and acts differently from the default style shown above. The list of custom stylesheets is currently limited to an approved set, but a mechanism is planned that will open this up to arbitrary stylesheets.

Other extensions are available at the repository and item levels, and new ones are in the works. It is even possible for individual repositories to specify custom extensions by defining them in 'Identify' response <description> elements, although this feature is

not fully developed yet. Having shown the promise of ERRoLs, though, a few words of caution are needed. ERRoLs operate by dynamically interacting with data providers via the OAI-PMH protocol. If these repositories are offline, slow, or less than fully OAI-compliant (which is frequently the case), the ERRoL functions will suffer. Nevertheless, these examples should show that ERRoLs are an interesting alternative to the conventional OAI model.

Ongoing Challenges for the OAI Community

We have highlighted a number of developments and ongoing work within the OAI community (and there are many more). But as the number of OAI data providers has grown, two broad areas of concern have arisen, particularly for service providers. These center on the variations and problems with data provider implementations and on the metadata itself. A third concern is the lack of communication among service and data providers. The metadata issues in particular have been well documented (Shreeves, Kaczmarek, & Cole 2003; Halbert 2003; Hagedorn 2003; Arms et al 2003), but we highlight some of the major issues in all areas of concern below.

Metadata Variation

While metadata must be created using unqualified Dublin Core (DC) encoding, as well as any other kind of encoding the data provider wishes, the choice of how to use the encoding standard and/or how to fit the encoding to metadata values that already exist varies widely among data providers. One institution's choice of how to use the DC Type tag can vary greatly from another's (e.g., "HTML" vs. "Preprint"). This can make it difficult to create a search environment in which users feel certain they are receiving

what they need. For instance, to normalize data (such as date or type elements) so search limiters can be used requires the development of common values among many disparate ones. The normalization of the subject element--with many different controlled vocabularies (or merely keywords) used by the different data providers--is, for most service providers, prohibitively resource intensive.

Metadata Formats

In the same vein, the problem of harvesting a data repository's additional metadata formats (beyond unqualified Dublin Core) can be a difficult task. For a large service provider with a standard method for processing harvested metadata, including new formats involves adding additional paths to the processing routines. The more formats, the more complex it becomes. Additionally, large service providers may have developed interfaces conforming to the simple Dublin Core standard and not have the ability to integrate more complex and more varied formats. For this, service providers need more all-encompassing game plans and better internal support.

OAI Data Provider Implementation Practices

The OAI protocol is flexible in that there are relatively few required pieces for implementation: valid responses to OAI verbs, the use of oai_dc, a unique and persistent OAI identifier, and a datestamp. The OAI Guidelines for Implementation have a limited technical scope, are intended for a general audience of implementers, and do not describe the consequences of not implementing some of the optional features of the protocol (Lagoze, Van de Sompel, Nelson, & Warner, 2002). This has meant that many of the

features of OAI, such as sets, use of descriptive containers, etc, that are quite helpful for service providers have been underutilized. Data providers need also to be aware of how their implementation of required items such as datestamps impacts service providers.

Communication issues

The OAI community is very loosely federated. There are general and technical listservs available through the Open Archives Initiative. However, as some of the issues above illustrate a serious need for best practices and guidelines exists for both data and service providers. An informal community of service providers has appeared who advise each other on the technicalities of performing harvesting and maintaining their service. While this ad-hoc community is welcome, a more formal method of communication between data and service provider is needed.

Future Directions

We have discussed above just some of the current developments in the OAI community. Below we outline some future directions. This list is not meant to be all-inclusive, but a taste of some of the ongoing research and practices in the OAI community.

Best Practices

As indicated above, service providers face serious challenges in both their harvesting and aggregating activities. The development of community specific best practices and implementation guidelines has been an important part of OLAC and other domain based

service providers. A group of service providers within the Digital Library Federation has now begun work on some more general best practices to be used with the DLF and beyond.

Static Repository Gateway

The technical hurdle is still sometimes too great for potential data providers. The Static Repository Gateway, developed at the Los Alamos National Laboratory, is the most recent option for OAI data providers and provides a very low entry point (Van de Sompel, Lagoze, Nelson, & Warner, 2004; Hochstenbach, Jerez, &Van de Somepl, 2003). Essentially, a resource developer can post a single large XML file containing the metadata and OAI wrappers on its webserver. This file can be accessed through an OAI gateway service. Currently two service providers, UIUC and the University of Michigan, have been working to shepherd potential data providers to one Gateway, which has proved very simple for both the service providers and data providers.

mod_oai Project

The mod_oai project, funded by the Andrew W. Mellon Foundation, is developing a tool that makes content that is accessible from Apache open-source Web servers available through the OAI protocol. This tool will essentially extend the benefits of selective and incremental harvesting available through the OAI protocol to the general web community (*mod_oai*, n.d.).

OAI-rights

The OAI-rights committee is working towards a means of incorporating structured rights statements about the resources exposed (i.e. the metadata) through the protocol (Lagoze, Van de Sompel, Nelson, & Warner, 2003). The committee does not intend to define a new rights language, but only to provide the means of communicating a structured, defined language within the protocol.

Controlled Vocabularies and OAI

Controlled vocabularies will become more important as data and service providers try to cope with the chaos that develops from aggregating metadata from diverse sources. Controlled vocabularies will become particularly important within self-archiving systems such as institutional repositories and e-print archives (many of which who are also OAI data providers); in many cases there is no cataloger to exert quality and authority control. A lightweight solution to this would be for authority agencies to mount their thesauri as an SRW/U search service, register it with the UIUC registry, and use ERRoLs to provide an HTML interface and URL access to items in the repository (*GSAFD Thesaurus*, n.d.).

SRW/U-to-OAI gateway to the ERRoL service

This service will allow institutions to load their data as an SRW/U search service, register it with the UIUC gateway, and automatically get OAI-PMH and ERRoL functionality for free. The OCLC OR Publications OAI repository is the first demonstration of this. This configuration adds searching capability to the mix of ERRoL features (*OCLC Research Publications*, n.d.)

References

Arms, W.Y., Dushay N., Fulker, D. & Lagoze, C. (2003). A case study in metadata harvesting: the NSDL. *Library Hi Tech 21*(2), 228-237.

Berners-Lee, T. (1998). *Hypertext Style: Cool URIs Don't Change*. Retrieved June 3, 2004, from http://www.w3.org/Provider/Style/URI.

Brogan, M. (2003). A Survey of Digital Library Aggregation Services. Washington, D.C.: Digital Library Federation. Retrieved June 2, 2004, from http://www.diglib.org/pubs/brogan/.

Davison, S., Requardt, C., & Brancolini, K. (2003). A Specialized Open Archives Initiative harvester for sheet music: a project report and examination of issues. Paper presented at the fourth International Conference on Music Information Retrieval: October 26-30, 2003, Baltimore, Maryland, USA. Retrieved June 2, 2004, from http://ismir2003.ismir.net/papers/Davison.PDF.

ERROLs for OAI Identifiers. (n.d.) Retrieved June 5, 2004, from http://www.oclc.org/research/projects/oairesolver/

Experimental OAI Registry at UIUC. (n.d.). Retrieved June 2, 2004, from http://oai.grainger.uiuc.edu/registry/.

Google Web APIs. (n.d.). Retrieved June 3, 2004, from http://www.google.com/apis/.

GSAFD Thesaurus (n.d.) Retrieved June 5, 2004, from http://errol.oclc.org/gsafd.oclc.org.html

Hagedorn, Kat. (2003). OAIster: a "no dead ends" OAI service provider. *Library Hi Tech 21*(2), 170-181.

Halbert, M. (2003). The Metascholar Initiative: AmericanSouth.Org and MetaArchive.Org. *Library Hi Tech 21*(2), 182-198.

Hochstenbach, P., Jerez, H., and Van de Sompel, H. (2003). The OAI-PMH Static Repository and Static Repository Gateway. In C.C. Marshall, G. Henry, & L. Delcambre (Eds) *Proceedings of the Third ACM/IEEE-CS Joint Conference on Digital Libraries: May 27-31, 2003, Houston, Texas* (pp.210-217). Los Alamitos, CA: IEEE Computer Society.

Lagoze, C., Hoehn, W., Millman, D., Arms, W. Gan, S. Hillmann, D., Ingram, C., Kraft, D., Marisa, R., Phipps, J., Saylor, J., Terrizzi, C., Allan, J., Guzman-Lara, S., and Kait. T. (2002). Core Services in the Architecture of the National Science Digital Library

(NSDL). In G. Marchionini & W.R. Hersh (Eds) *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries: July 14-18, 2002, Portland, Oregon* (pp. 201-209). New York: ACM Press.

Lagoze, C. & Van de Sompel, H. (2001). The Open Archives Initiative: building a low-barrier interoperability framework. In E.A. Fox & C.L. Borgman (Eds), *Proceedings of First ACM/IEEE-CS Joint Conference on Digital Libraries: June 24-28, 2001: Roanoke, Virginia, USA* (pp.54-62). New York: ACM Press.

Lagoze, C. & Van de Sompel, H. (2002). *The Open Archives Initiative Protocol for Metadata Harvesting – version 2.0*. Retrieved June 2, 2004, from http://www.openarchives.org/OAI_protocol/openarchivesprotocol.html.

Lagoze, C. & Van de Sompel, H. (2003). The making of the Open Archives Initiative Protocol for Metadata Harvesting. *Library Hi Tech 21*, 118-128.

Lagoze, C., Van de Sompel, H., Nelson, M., & Warner, S. (2002). *Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting*. Retrieved June 2, 2004, from http://www.openarchives.org/OAI/2.0/guidelines.htm.

Lagoze, C., Van de Sompel, H., Nelson, M., & Warner, S. (2003). *OAI-Rights White Paper.* Retrieved June 2, 2004, from http://www.openarchives.org/documents/OAIRightsWhitePaper.html.

Language Data Consortium. (n.d.). Retrieved June 2, 2004, from http://wave.ldc.upenn.edu/.

Linguist List. (n.d.) Retrieved June 2, 2004, from http://cf.linguistlist.org/.

McKiernan, G. (2003a). Open archives initiative service providers. Part I: science and technology. *Library Hi Tech News 9*.

McKiernan, G. (2003b). E-profile: Open archives initiative service providers. Part II: social sciences and humanities. *Library Hi Tech News 10*.

McKiernan, G. (2004). E-profile: Open archives initiative service providers. Part III: general. *Library Hi Tech News 1*, 38-46.

mod_oai. (n.d.). Retrieved June 3, 2004, from http://www.modoai.org/.

National Science Digital Library. (n.d.) Retrieved June 2, 2004, from http://www.nsdl.org/.

OCLC Research Publications. (n.d.) Retrieved June 5, 2004, from http://errol.oclc.org/orpubs.oclc.org.html

Open Archives Initiative. (n.d.). Retrieved June 2, 2004, from http://www.openarchives.org/.

Open Archives Initiative Registered Data Providers. (n.d.) Retrieved June 3, 2004, from http://www.openarchives.org/Register/BrowseSites.pl

Open Language Archives Community. (n.d.). Retrieved June 2, 2004, from http://www.language-archives.org/.

Open Language Archives Community Participating Archives. (n.d.) Retrieved June 3, 2004, from http://www.language-archives.org/archives.php4.

Sheet Music Consortium. (n.d.). Retrieved June 2, 2004, from http://digital.library.ucla.edu/sheetmusic/.

Sherman, C. & Price, G. (2003). The invisible web: uncovering sources search engines can't see. *Library Trends 52* (2), 282-298.

Shreeves, S.L., Kaczmarek, J.S., & Cole, T.W. (2003). Harvesting cultural heritage metadata using the OAI protocol. *Library Hi Tech 21*(2), 159-169.

Simons, G. & Bird, S. (2003). Building an Open Language Archives Community on the OAI foundation. *Library Hi Tech 21* (2), 210-218.

SRW- Search/Retrieve Web Service. (n.d.). Retrieved June 3, 2004, from http://www.loc.gov/z3950/agency/zing/srw/.

Van de Sompel, H., Lagoze, C., Nelson, M., and Warner, S. (2002). *Implementation Guidelines for the Open Archives Initiative for Metadata Harvesting: The OAI Static Repository and Static Repository Gateway*. Retrieved Sept. 3, 2004, from http://www.openarchives.org/OAI/2.0/guidelines-static-repository.htm.

Van de Sompel, H., Young, J.A., & Hickey, T.B. (2003). Using the OAI-PMH… differently. *D-Lib Magazine 9* (7/8). Retrieved June 3, 2004, from http://www.dlib.org/dlib/july03/young/07young.html.

Warner, S. (2003). E-prints and the Open Archives Initiative. *Library Hi Tech 21* (2), 151-158.

Waters, D. (2001). The metadata harvesting initiative of the Mellon Foundation. *ARL Bimonthly Report 217.* Retrieved June 3, 2004, from http://www.arl.org/newsltr/217/waters.

ZeeRex. (n.d.). Retrieved: June 3, 2004, from http://explain.z3950.org/.

Zia, L.L. (2001). Growing a national learning environments and resources network for science, mathematics, engineering, and technology education. *D-Lib Magazine 7* (3). Retrieved June 2, 2004, from http://www.dlib.org/dlib/march01/zia/03zia.html.